图书情报工作
第63卷 第20期 2019年10月

基于成果特征的学者学术专长识别方法

■ 陈翀[1] 李楠[1] 梁冰[2] 王晨琳[1] 徐曾旭林[1] 郑婷婷[1]
[1] 北京师范大学政府管理学院 北京 100875 [2] 中国科学技术信息研究所 北京 100038

摘要：［目的/意义］基于成果特征标识学者的学术专长是学者画像的重要任务，对学者分类、评审专家遴选、发现小同行等应用具有重要价值。［方法/过程］首先分析揭示学术专长的因素，用层次分析法构造专长标签权重分配模型；采用 TextRank 和概念链接技术从中英文成果内容中识别主题术语，结合权重筛选出具有领域共识和专长概括性的词汇作为专长标签。选取获得人才称号的多个领域科研人员，从中文或英文代表成果中提取专长标签，以人才公示中的专长领域作为对照基准，通过人工打分和语义计算评测识别效果。［结果/结论］在被贴中文专长标签的学者中，71.9%的个体的专长描述被认为满意。在被贴英文专长标签的学者中，77.2%的个体的专长描述被认为满意。实验表明提出的学者学术专长识别方法具有合理性。主要创新在于：在中英文不同语种以及是否存在外部知识库的条件下，提出从文献内容中挖掘候选标签词的解决方案；结合计量因素，用多种成果特征筛选专长标签，并提出权重分配的方案；针对评价基准欠缺的问题，提出基于语义计算的方式补充答案，从而扩充评价手段。

关键词：学者画像 专长标签 层次分析法 术语提取 专长标识评价方法
分类号：G250.7
DOI:10.13266/j.issn.0252-3116.2019.20.011

1 引言

在人类社会迈向知识时代的进程中，掌握知识的人成为越来越有价值的资源。学者是这类资源中的典型，具有丰富特征[1]，其学术专长对标识学者知识特征最为重要，可用于对学者的分类、检索，帮助发现小同行[2]以便促进交流合作，还可用于更准确地遴选论文或项目的评审专家[3-4]。同时，文献检索系统、个性化学习、知识协作等面向知识群体的应用也需要根据用户的学术专长改进服务精准度。因而，标识学者的学术专长是一个极具现实意义的研究问题。

在本文中，学术专长是指学者擅长的研究方向。基于成果特征刻画学者学术专长的核心问题有两方面：一是分析体现学者专长的因素；二是从成果中识别具有领域共识且概括专长的词汇作为标签。为了表述方便，笔者将成果内容中概括性强、规范性好的词汇称为主题术语；能标识学者专长的词汇称为专长标签，主题术语为专长标签提供候选词。

目前，获取学者专长的方法有两类：一是从个人主页、简历等来源直接提取；二是从学者的成果内容中识别恰当的词汇。在前一种方式中，学者给出的专长描述词汇数量少、表述习惯因人而异，且研究表明仅有21.3%的学者会在其主页中给出研究兴趣[5]，所以提取结果的完善度、规范性和及时性并不能得到保证。现有工作通常从学者发表的论文等成果中提取其研究兴趣[1]。学者成果内容扩展了专长标签的来源和数量，可使标签更客观全面，还可从学者新发表的成果中及时发现新的兴趣标签。但从成果中自动识别出的词汇如何能兼具领域共识和专长代表性，是值得深入研究的问题；特别是当学者群体不限于特定领域，且成果有中英文不同语种时，这一问题就更具有挑战性。

笔者首先分析学者成果中能揭示其学术专长的重要因素，如学者对成果的贡献、成果对学术界的贡献以及用于概括成果的术语质量等。其次，采用 TextRank 和概念链接技术，分别识别中文和英文成果中的主题术语。最后根据层次分析法构造权重分配模型，按照权重对候选词打分，筛选出能反映学者研究专长的标

作者简介：陈翀(ORCID:0000-0002-9704-1575)，副教授，博士，E-mail:chenchong@bnu.edu.cn；李楠(ORCID:0000-0002-5724-8926)，硕士研究生；梁冰(ORCID:0000-0002-7622-6618)，高级工程师，博士；王晨琳(ORCID:0000-0002-0640-9339)，本科生；徐曾旭林(ORCID:0000-0002-8181-1168)，本科生；郑婷婷(ORCID:0000-0003-4542-7908)，硕士研究生。

收稿日期：2019-01-25 修回日期：2019-06-03 本文起止页码：96-103 本文责任编辑：徐健

签,从而解决学术专长识别问题。

2 相关研究

刻画学者的学术专长属于学者画像的一个重要方面。学者画像是通过分析学者的个人描述信息、成果或学术行为,识别并提取恰当的标签来概括学者的个人特征、研究兴趣以及学术影响力[1]等。信息环境中用户画像的目的是为用户提供更加精准的个性化服务[6],而学者画像可用于改进对科研群体的精准化和个性化服务,识别学者专长就是其中的核心。

2.1 专长特征分析

能够反映学者专长的信息包括学者发表的成果[7]、承担的科研项目[8]、合作关系[9]、论文引用[10]等。其中,有两类信息对揭示学者的专长非常重要,即成果内容特征和学者对成果的贡献度。前者是指对成果内容有重要概括度且有领域共识的词汇,后者体现学者与成果的关联以及成果对学术界的价值。

2.1.1 成果内容特征　在论文、项目等成果中,作者给出的关键词与成果的研究主题密切相关,但这些词侧重于标引成果本身,而不是侧重于表达作者的专长。而且,作者关键词普遍存在用词不规范现象,例如语义粒度不均匀[11]、标引深度把握不当以及通用词标引过多[12-13]等。因此,直接用作者关键词标识学者专长并不理想,有时还会引入噪音词汇。鉴于成果内容中包含更加丰富的信息,从中挖掘出有重要概括度且有领域共识的词汇来补充学者专长描述是一种更好的思路。相关研究通常基于论文的标题、摘要[14]、段首及结尾内容[15]等部分识别重要词汇。

2.1.2 学者对成果的贡献度　当今社会中的科研合作现象普遍存在。统计表明 2015 年国内科技论文领域的合著论文数量占论文总数的 92.3%[16]。对图情领域 4 种核心期刊的分析发现,2－3 人合著发文成为主流,而 4 人及 4 人以上合著将成为未来合著的趋势[17]。每个合作者对成果的贡献并不相同,这就意味着同一成果在揭示每位作者专长上的价值并不等同。因此,需要选择代表性成果来体现作者的实质性贡献。在科研计量评价中,一些研究常常会经验性地按照作者署名位序区分作者贡献[18]。问卷调查显示,约 82% 的被调查者认为署名顺序与作者贡献相关,署名越靠前的作者对研究成果贡献越大[19]。根据多数学科领域的署名习惯,通常认为第一作者和通讯作者与成果贡献的关系更密切[20]。此外,由于学者在其研究生涯中一般会有多篇发文,发文量[21]、发文时间、被引量[22]等因素对确定学者对成果的贡献也有参考价值。学者的发文量和发文时间可以反映其领域活跃度,而被引量则可以反映其学术成果的质量以及同行的认可程度。在对学者影响力排名中,综合考虑上述重要的影响因素构建指标被认为更具区分效果、更全面精细[21,23]。因此,笔者选用成果被引量、作者署名位序因素筛选学者专长标签。

2.2 基于成果内容的主题术语识别

从成果内容中识别学者专长标签的一个基本任务是发现成果中具有领域共识的主题术语。按照是否依赖于外部知识可大致分为直接和间接两种方式。直接方式的优点是不依赖于已有的领域知识体系;间接方式是利用词典、本体、知识图谱等领域知识库,将成果内容通过语义计算映射到规范受控的领域术语空间中,其优点是使标签词汇在语义和粒度上更规范。

2.2.1 直接方式　此类方法重点在于度量词汇对整体内容的重要性。从实现上,可以通过构建词汇共现网络,用 TextRank 等算法识别网络中重要度高的节点[24-26]作为反映学者专长的关键词汇;还可以考虑词汇在文档中的位置影响力[25]、在网络中的覆盖能力[27],或结合词频、词语位置等特征衡量其重要性。除了成果本身,借助与其相关的人及行为也能发现揭示内容的重要词汇,例如借鉴从博客内容中筛选标签的做法,考虑评论、引用、链接等外在特征计算内容中关键词的重要度[28]。目前使用较多的还有基于语义计算的方法,例如采用主题模型得到成果中蕴含的主题分布以及主题对应的词汇分布,并将出现概率大的主题及词汇作为标识学者专长的依据[29,30]。但这样的结果对人类的可解释性并不友好,因为主题的语义是通过若干词的分布隐含地表达,而且概率大的主题仅表示论文中频繁出现的主题,不一定能揭示学者的领域专长。词嵌入表示方式提出以后,有研究采用 Word2vec 构造词汇向量来改进语义计算的灵活性,将词汇共现网络中的节点用词向量表示,通过计算词汇相似度来改进候选关键词的权重分配,提升领域关键词识别准确性[31]。

2.2.2 间接方式　该方法重点是选取成果中的重要词汇并向规范受控的领域术语空间进行合理映射。外部知识库包括基于期刊文献构造的专长词典[32],以及现有的领域概念本体[33]等。借助外部知识库进行重要概念识别通常采用概念链接技术[34-36]。以维基百科为例,具体做法是将维基百科文章页面的标题词汇作为概念术语,先用词性、词频等一般的自然语言处理

手段识别成果内容中的重要词汇;用词汇在成果内容和维基百科中出现的统计信息以及词汇之间的关联关系,将其自动映射到维基百科概念术语,从而实现成果中词汇的规范化。这类方法不但可以用于术语规范化,还可以用于语义消歧等诸多任务,主要工具有 Tagme[37]、Wikifier[38]等。

2.3 本文与相关研究的不同之处

在识别学者学术专长上,有一些前人研究与本文的任务相关。刘晓豫等[13]以规范处理后的论文关键词为特征,基于重叠 K-Means 聚类算法识别大数据领域的专家专长类别,但该研究并未探讨如何自动形成能恰当标识专家专长的标签;毛进等[39]以论文全文中的高频名词为特征,构建计算语言学领域专家图谱,但该研究没有考虑到同一篇论文的合作者对成果的贡献不尽相同,同样的专长词汇来标识不同贡献的人并不恰当;范晓玉等[40]基于成果内容和计量特征提取科研偏好标签,并根据作者贡献和成果新旧来调整标签权重,但该研究只针对选定的两位学者进行实现,并未对大量学者的专长标签识别进行方法和效果上的检验。

综上,本研究与现有工作的不同之处主要有:①对中英文不同成果,提出从其内容中自动识别具有领域共识、且粒度合理的主题术语,作为候选的专长标签。相比于从简历等来源直接抽取,该方法更加客观,保证了标签的丰富性与多样性,而且能及时发现学者新的专长标签。②将学者对成果的贡献度作为候选标签的权重因素,有助于在科研合作普遍的情况下,区分不同作者的专长差别。③采用层次分析法筛选专长标签,将定性与定量方式相结合,不仅计算简便,而且结果可解释性更好,在实际应用中更容易被用户理解。④面对缺乏标准评测集的现状,采用人工打分与语义计算相结合的实验设计与评价方法,最大程度地补充了专长标签的评价手段。

3 专长标签识别方法

从成果中识别专长标签的朴素假设是:学者近期代表作对体现其学术专长有重要价值。可以从学者贡献较大的近期重要成果中找到有领域共识的规范词汇来筛选专长标签。笔者采用层次分析法对体现学者专长的重要因素构造权重分配模型,通过权重分析确定各项特征的重要程度,然后用直接及间接的方式计算学者成果中所有候选主题术语的分值,并综合以上因素筛选高分词汇作为专长标签。在从成果内容中获取主题术语时,由于开放的多领域优质中文知识库比较欠缺,而英文知识库则有公认质量较高且覆盖领域广泛的维基百科,因此笔者对中英文分别采用了直接和间接的方式。

3.1 标签权重分配模型

设学者 r 的成果集合为 D,专长标签集合为 A。标识 r 的专长标签应满足:①能反映 r 研究领域特征,具有概括性和规范性,用 B_1 表示;②来自 r 有实质贡献的重要成果,即体现 r 对该专长领域的贡献度,用 B_2 表示。

根据 2.1 节对专长特征的调研,定义如图 1 所示的学者专长标签权重分配模型。B_1 对应的因素主要包括原文关键词 C_1 和候选主题术语 C_2,B_2 对应的因素主要包括署名位序 C_2 和成果被引量 C_4,它们分别表示 r 对成果的贡献,以及学术群体对成果的认可程度。

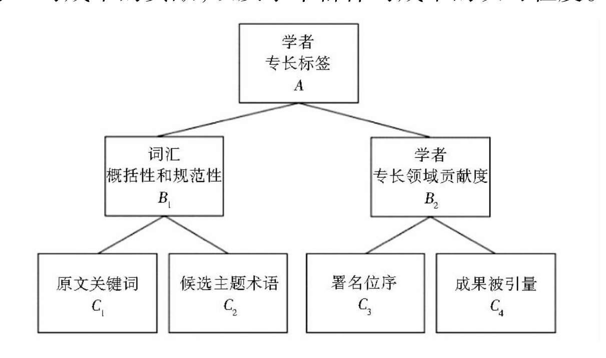


图 1 学者专长标签权重分配层次模型

笔者用层次分析法计算专长标签的特征权重,得到描述学者专长的标签。层次分析法是一种定量定性相结合的方法,将决策问题分解为不同目标的层次结构,通过求解判断矩阵特征向量,找到每一层次的各元素对上一层次某元素的优先权重,再逐层归并得到总目标的最终权重,以最大的最终权重作为最优方案。该方法的结果有一定可解释性,在学者学术影响力评价[41]、科技人员论文学术价值评估[42]等很多方面均有应用。

采用德尔菲法对图 1 确定各因素两两间相对重要程度,构造判断矩阵 M。M_{ij}代表某上层因素对应的下层因素 i 与 j 的重要性比较结果,如公式(1)所示:

$$M_{ii}=1, M_{ij}=\frac{1}{M_{ji}} \quad (i,j=1,2) \qquad \text{公式(1)}$$

对于因素 B_1 和 B_2,与词汇的概括性和规范性相比,r 对其专长领域的贡献度对识别专长标签更有影响,认为 B_2 比 B_1 稍微重要,因此确定 $A \rightarrow B$ 判断矩阵 $M_0=\begin{bmatrix}1 & 1/2\\ 2 & 1\end{bmatrix}$。

陈翀，李楠，梁冰，等. 基于成果特征的学者学术专长识别方法[J]. 图书情报工作,2019,63(20):96－103.

对于因素 C_1 和 C_2，在科研成果中，原文关键词 C_1 是作者给出的，在表达成果内容和主题上有天然的重要性，很多研究也会用关键词进行主题挖掘和专长识别。C_2 是从内容中自动识别的词汇。相比之下，C_1 的受认可程度比 C_2 更大，即 C_1 比 C_2 稍微重要，因此确定 $B_1 \rightarrow C$ 判断矩阵 $M_1 = \begin{bmatrix} 1 & 2 \\ 1/2 & 1 \end{bmatrix}$。

对于因素 C_3 和 C_4，由于科研合著发文的普遍性，C_3 是成果作者共同认可的署名次序，能体现 r 对成果的实质贡献程度，C_4 是成果本身的领域贡献度。即使 r 的成果被引量高，但如果其署名位序靠后且不是通讯作者则不能说明其贡献大。在体现 r 对成果实质贡献上，C_3 比 C_4 明显重要，因此确定 $B_2 \rightarrow C$ 判断矩阵 $M_2 = \begin{bmatrix} 1 & 4 \\ 1/4 & 1 \end{bmatrix}$。

以上 3 个判断矩阵均通过一致性检验，说明矩阵中各因素相对重要性不存在自相矛盾。最后，求得 M 的特征向量 $W = (0.22, 0.11, 0.54, 0,13)$。各级指标对应因素的权重系数如表 1 所示：

表 1　各级指标及其权重表

总目标	一级指标	权重	二级指标	权重
学者专长标签 A	词汇概括性和规范性 B_1	0.33	原文关键词 C_1	0.22
			候选主题术语 C_2	0.11
	学者专长领域贡献度 B_2	0.67	署名位序 C_3	0.54
			成果被引量 C_4	0.13

3.2　主题术语识别

对于任一学者 r 的成果文档集合 $D = \{d_1, d_2, \cdots\cdots, d_n\}$，通过直接或间接方式识别出的领域主题术语集合记为 $T_D = \{t_1, t_1, \cdots\cdots, t_m\}$。本小节要实现的任务是 $D \rightarrow T_D$，即从成果内容中识别主题术语。

3.2.1　中文主题术语识别　由于缺少开放的多领域中文知识库，笔者采用直接方式生成 T_D。首先对 D 构造词汇共现网络，计算词汇节点在网络中的重要度。TextRank[24] 是在词汇共现网络中衡量节点重要性的常见方法，利用投票原理，在给定共现窗口阈值内，根据节点间的连接关系获得每个节点对邻居节点的投票，票的权重取决于自己所得的票数。某节点的 TextRank 值由邻居节点投票计算得到，依据 TextRank 值排序可得到候选主题术语。

3.2.2　英文主题术语识别　采用间接方式从英文成果中识别 T_D。以维基百科页面标题作为概念，对成果内容中的重要词汇用概念链接技术计算出映射到概念的链接概率和一致性[37]，从而得到重要词汇的受控概念。这一做法避免了词汇存在同义和二义现象对标识专长造成的影响。图 2 为 Tagme 标注工具对某英文摘要进行主题术语识别的结果示例。该方法将文中"machine learning techniques"映射到维基百科"Machine Learning"，这一较规范的词汇可以作为候选主题术语。

图 2　Tagme 概念链接工具在英文摘要中标注术语的结果示例

3.3　专长标签选择

在图 1 中，从 C_1、C_2 到 B_1 的阶段是在学者 r 的主题术语集合 $T_D = \{t_1, t_1, \cdots\cdots, t_m\}$ 和成果的作者关键词集合 T_k 中选择出专长标签集合 T'_D，实现任务 $(T_D, T_k) \rightarrow T'_D$。计算 r 的候选主题术语集合 T_D 中 $\forall t \in T_D$ 对应的权重，若 $t \in T_k$，则 $t_{C_1} = 1$ 且 $t_{C_2} = 0$，反之则 $t_{C_1} = 0$ 且 $t_{C_2} = 1$。

从 C_3、C_4 到 B_2 的阶段：用 m 表示作者署名位序，令 $t_{C_3} = 1/m$，将通讯作者视为第一作者位序。用 n 表示 t 所在成果被引量，n' 表示 r 所有成果被引量之和，则 $t_{C_4} = n/n'$。

最后，通过公式(2)计算每个主题术语 t 作为候选专长标签的得分 $s(t)$。比较同一学者 r 的所有成果的候选专长标签得分，并将相同标签对应得分求和。最后，选择得分大于特定阈值或指定数量的标签作为 r 的专长标签。本文实验中选择了得分前 5 位的标签。

$$s(t) = 0.22 \times t_{C_1} + 0.11 \times t_{C_2} + 0.54 \times t_{C_3} + 0.13 \times t_{C_4} \quad \text{公式(2)}$$

4 实验及结果

4.1 数据与方法

4.1.1 实验数据 笔者从北京市科学技术委员会官网公布的"北京市科技新星"人才计划名单,选取 2013－2017 年入选的科研人员 557 人,专业涵盖理工农林医等各个领域。通过百度学术及个人主页等来源获取上述人员在近 5 年的论文和项目成果数据,包括标题、关键词及摘要等。对同一作者的论文按被引量降序排列,在前 5 篇中,如果中文占多数就选 3 篇中文论文作为其代表作,反之选 3 篇英文论文。如果可获取的成果不足 5 篇就取多数语种中的 2 篇作为代表作。剔除数据缺失值较多的人员,共得到科研人员 480 人,记为 R。其中对应中文成果的人才 300 人,记为 R_c,中文成果数为 867 篇;对应英文成果的人才 180 人,记为 R_e,英文成果数为 520 篇。

4.1.2 识别方法 对中文成果的标题、摘要等数据经过 HanLP 中文处理工具包预处理然后用 TextRank 算法结合词频和词长等启发式规则提取论文中重要的词汇作为主题术语。对英文成果用 Tagme 工具将论文摘要中的词汇映射到维基百科概念。得到候选主题术语后,根据层次分析法计算得分最高的 5 个词汇作为学者的专长标签。

4.2 实验结果

表 2 和表 3 中分别列出了基于中文和英文成果的科研人员专长标签识别结果样例。其中,S 为"科技新星"申报者 r 的专业领域,在本文中作为评测基准。T_k 为申报者近年影响力高的代表作中的作者关键词集合,T_D 为从代表作中识别出的候选主题术语集合,T'_D 为本文方法得到的专长标签词汇集合。直观上,从内容中提取出的主题术语 T_D 与 T_k 有一定的重合度,但语义粒度不那么细微,能表达更大的主题,而 T_k 的词汇粒度较不均匀,例如 r_1 的 T_k 中"硬膜外麻醉"比 T_D 中"患者麻醉"粒度更细,而 r_4 的 T_k 中"Quality of service(服务质量)"又过于笼统,粒度太粗。

表 2 中文专长标签识别结果样例

r	S	T_k	T_D	T'_D
r_1	临床麻醉	麻醉;异氟醚;异丙酚;丙泊酚;剂量;全麻;硬膜外麻醉……	丙泊酚;丙泊酚剂量;异丙酚;异氟醚;结肠癌;患者麻醉……	丙泊酚;异丙酚;异氟醚;丙泊酚剂量;患者麻醉
r_2	城市轨道交通	城市轨道交通;轨道交通产业构成;全生命周期;综合管理……	轨道交通;城市轨道;生命周期;智能管理;综合管理……	城市轨道交通;轨道交通;轨道交通产业链;全生命周期;综合管理

表 3 英文专长标签识别结果样例

r	S	T_k	T_D	T'_D
r_3	神经外科	Ganglioglioma(神经节细胞胶质瘤);Intraventricular(脑室内);Hydrocephalus(脑积水);Prognosis(预后);Von Hippel-Lindau disease(VHL 综合征);Hemangioblastoma(血管母细胞瘤)……	Ganglioglioma(神经节细胞胶质瘤);Hydrocephalus(脑积水);Ventricular system(脑室系统);Surgical pathology(外科病理学);Von Hippel-Lindau disease(VHL 综合征)……	Ganglioglioma(神经节神经胶质瘤);Hydrocephalus(脑积水);Ventricular system(脑室系统);Von Hippel-Lindau disease(VHL 综合征)Surgical pathology(外科病理学)
r_4	计算机软件	Quality of service(服务质量);service-oriented architecture(面向服务的体系结构);Computational modeling(计算建模);Big Data(大数据);Web services(Web 服务);Petri nets(Petri 网)……	Basel problem(巴塞尔问题);Cloud computing(云计算);service-oriented architecture(面向服务的体系结构);Interval arithmetic(区间算法);Web services(Web 服务);Petri nets(Petri 网)……	service-oriented architecture(面向服务的体系结构);Big Data(大数据);Web services(Web 服务);Petri nets(Petri 网)Computational modeling(计算建模)

5 评测与分析

在与本文同类的研究中,一个最突出的困难是没有公认的标准评测集。笔者选用科技新星人才公示的专业领域作为评测集 S。但 S 中的词汇数量少、粒度粗,直接作为评测标准并不理想。因此评判依据分为两部分:一是来自人工打分;二是采用词向量相似度判断 T'_D 与 S 的语义关联。为了使人工评测结果具有可信性,首先对评测者进行领域辨识能力检测,笔者招募了来自工科、理科、社科等不同专业的 10 名志愿者。志愿者通过检测并进行标注训练,才能对科研人员的专长标签进行正式打分,最后经评分一致性检验保证人工打分的客观性。

5.1 评测者领域辨识能力检测

领域辨识能力实验主要是分析志愿者对各专业领域的了解程度和判断领域专长的能力。检测方法:在 R_c 和 R_e 中各随机选择 50 名科研人员的成果摘要和关键词,并以选择题形式展示包含标准答案在内的 4 个相似领域词汇;志愿者根据对成果内容的理解,选择最相关的专业领域;根据正确率评价志愿者的辨识能力。选项中与标准答案相似的领域词汇均是通过 Word2vec

计算词向量相似度得到的。在表4中，根据科技新星人才公示，特定 r 的专业领域是“化学电源”，它是标准答案；用词向量算出S中与其语义最相近的其他3个词汇选项；如果志愿者选择C就记作正确，否则记作错误。

表4　领域辨识能力实验题目示例

题目描述	请阅读下面的论文摘要和关键词，选择与之最相关的作者研究领域。
题目内容	摘要：根据电池的外观、电性能、环境适应性和安全性等检测项目，鉴于部分检测过程中可能存在的火爆炸、漏气漏液、噪声振动、机械和电气等危险，从人员、样品、设备和环境等方面，提出了相应的安全防护要求和建议。 关键词：电池 检测 实验室 安全防护
题目选项	A 化学 B 材料化学 C 化学电源 D 环境化学

在中英文领域判断实验中，正确率达到90%的志愿者分别为9人和4人。因此，他们可以作为评测者分别对生成的中英文专长标签进行人工打分。

5.2　专长标签人工评价

人工评价主要是让评测者根据领域知识和评测标准判断特定学者 r 的专长标签的合理性。对 $\forall r \in R_c$，提供 r 的中文专长标签及其中文代表作；对 $\forall r \in R_e$，提供 r 的英文专长标签及其英文代表作。打分结果分为满意(1)、不确定(0)或者不满意(-1)。“满意”表示反映专长、语义粒度合理且词汇具有领域共识；“不满意”表示不反映专长、语义粒度不合理或无领域特色。每位评测者需独立完成 R_c 或 R_e 中所有学者专长标签集合的打分。

采用SPSS软件并计算Kendall’s W系数对评测者的评分进行一致性检验。最终得到中文实验评测者的评分一致性系数0.924，英文实验评测者的评分一致性系数0.921，在检验水平为0.05时，P值均小于0.001，检验结果显著。结果表明评测者对标签合理性打分判断具有很强的一致性。

按照服从多数的原则，在每个 r 的标签打分结果中，如果评分为“1”的数量达到评分者半数及以上，则认为对 r 自动生成的标签满意。经统计，R_c 及 R_e 中被判断为专长标签满意的科研人员占比 P_1 分别为65.3%和68.3%。打分结果表明，笔者从中英文论文中识别出的学者专长标签具有一定的合理性。

5.3　专长标签语义计算补充评价

由于评测标准词数量少、语义粒度粗，不足以给评测人员提供更丰富的领域信息。为此对判定结果不为“1”的学者集合 R'，$R'=R'_c \cup R'_e$，用语义计算对其专长标签作补充评价。训练Word2vec词向量，采用余弦公式计算 R' 对应的评测标准词 S 及专长标签 T'_D 的语义相似度 sim，相似度超过特定阈值的标签也被认为满意。对 $\forall r \in R'_c$ 或 R'_e，计算其专长标签与评测标准的词向量相似度，如果 $sim \geq 0.9$ 的专长标签数达到半数及以上，则认为对 r 生成的标签可达到满意。在 R'_c 及 R'_e 中，专长标签满意的科研人员比率 P_2 分别为6.6%和8.9%。

词向量生成方法具体如下：将所有中英文成果的标题、关键词、摘要以及对应学者专业领域分别进行文本预处理得到分词元素，而后调用gensim的Word2vec模块训练并生成词向量模型，每个分词元素对应一个100维的词向量。其中，预处理包括切分、标注词性、去除停用词，保留名词、动词、形容词等词性。任一词汇的词向量是由分词后每个元素的词向量进行平均池化得到。

5.4　实验结果分析

经过上述两阶段评判，中文和英文专长标签满意的科研人员总比率 P 分别达到了71.9%和77.2%。考虑到评价标准S在数量和语义粒度上的不足，这样的标识正确率能够说明笔者所提方法给出的学术专长标签的合理性。本文方法突破了文献关键词的粒度不一、细粒度词汇多的局限，可以用具有概括性的恰当词汇来表示学者专长；而作为评价标准的专长描述来自人才公示，这一信息通常粒度过粗，例如 r_1 的评价基准是“临床麻醉”，所以无法从字面上确认匹配情况，笔者采用语义吻合或接近的原则进行评价，结果见表5。在人工打分实验中，由于一些科研人员的领域专业性较强，对评测者形成一定程度的认知障碍，而语义计算实验减少了主观判断失误造成的影响。

表5　专长标签满意的科研人员比率分析

评价指标	人工打分判定标签满意的人员比率 P_1	语义计算判定标签满意的人员比率 P_2	标签满意的人员总比率 P
中文标签	65.3%	6.6%	71.9%
英文标签	68.3%	8.9%	77.2%

本研究也注意到文本语料的规模限制了词向量的计算，导致部分领域词汇以及专长标签无法计算词向量。总体而言，笔者所提方法能较好地识别出用来标识学者专长的中英文标签，并且在主题术语识别、标签权重分配和评价方法上具有一般性，能够推广到不同领域的学者专长标识。

6　结论和展望

笔者分析学者成果中揭示其学术专长的重要因

素,构造层次模型得到权重分配;采用 TextRank 和概念链接技术分别识别中英文成果内容中的主题术语,为描述专长提供更丰富的合理候选词;最后根据多种权重因素筛选具有领域共识和专长概括性的标签,从而解决学者专长识别问题。以人工打分和语义计算实验评价专长标签识别效果,表明笔者所提方法具有一定的合理性。在测试数据上,中英文标签结果满意的人员占比分别为 71.9% 和 77.2% 。

本文的主要贡献在于:①提出并实现了特定领域大规模识别学者专长标签的方法。综合学者的成果内容及学术贡献等多种特征建立层次分析模型,挖掘描述学者专长知识的主题术语,进而识别出粒度均匀且具有领域共识的学者专长标签。②探索了主题术语直接和间接生成方法。针对中英文不同语种及有无外部知识库的情况提出解决途径,拓宽了本文的应用范围。③提出了合理的实验设计与评价方案。将人工打分与语义计算相结合来评价专长标签识别方法。针对评测标准词汇数量少、语义有限造成人工评价依据不足的情况,采取了语义计算的补充策略解决评价受限的问题。

未来研究的改进之处如下:①专长标签识别实验数据有限。后续拟在更大的中文和英文数据集上对学者专长标签识别效果进行实验评测。②实验设计有待深入细化。在对成果特征用层次分析法进行建模时,判断矩阵的赋值有一定主观性,且一些领域并非按照作者贡献来署名。后续研究可融入同行评议意见、作者贡献说明等因素优化权重分配,使结果更加合理。人工评测粗略判断结果的整体合理性后,应该进一步具体判断每个标签的识别效果。在语义计算评价时,在更大的领域数据集上训练词向量模型将有助于改进词汇表示的合理性。③中英文文本混合情况下通用识别方法的探索。在实际数据中,中文成果也有英文术语存在,不能完全将其按照某一种语种进行区分,应采用更加通用的方法来识别学者的专长标签。

参考文献:

[1] 袁莎,唐杰,顾晓韬. 开放互联网中的学者画像技术综述[J]. 计算机研究与发展,2018,55(9):1903 - 1919.
[2] 朱伟珠,李春发. 基于概念知识网络的"小同行"评议专家遴选方法实证研究[J]. 情报杂志,2017,36(7):78 - 83,88.
[3] 赵丽莹,冯树民,刘彤,等. 如何选择"小同行"审稿专家[J]. 编辑学报,2007,19(1):75.
[4] 程薛柯. 科技项目小同行评审专家识别研究[D]. 北京:中国科学技术信息研究所,2016.
[5] TANG J, YAO L, ZHANG D, et al. A combination approach to web user profiling[J]. ACM transactions on knowledge discovery from data (TKDD), New York: ACM, 2010,5(1):1 - 44.
[6] 胡媛,毛宁. 基于用户画像的数字图书馆知识社区用户模型构建[J]. 图书馆理论与实践, 2017 (4):82 - 85.
[7] 巩军,刘鲁. 基于个人知识地图的专家推荐[J]. 管理学报,2011,8(9):1365 - 1371.
[8] TANG J, ZHANG D, YAO L. Social network extraction of academic researchers[C]// Seventh IEEE international conference on data mining. Piscataway: IEEE, 2007:292 - 301.
[9] YAN M, YU Z, ZHANG Y, et al. An expert recommendation approach combining project correlation and professional ability[C]// International conference on fuzzy systems and knowledge discovery. Piscataway: IEEE, 2015:1220 - 1224.
[10] 张思凤,梁梦丽,曹高辉. 基于引文的科技文献主题抽取研究[J]. 情报理论与实践,2017, 40(6):122 - 127.
[11] 邓启明,王景辉. 关键词标引中常见问题与分析[J]. 科技与出版,1999(2):36.
[12] 王思哲. 我国学术期刊关键词标引质量探析[J]. 延安大学学报(社会科学版), 2001,23(3):97 - 99.
[13] 刘晓豫,朱东华,汪雪锋,等. 多专长专家识别方法研究——以大数据领域为例[J]. 图书情报工作,2018,62(3):55 - 63.
[14] 任海英,王德营,王菲菲. 主题词组合新颖性与论文学术影响力的关系研究[J]. 图书情报工作,2017,61(9):87 - 93.
[15] 赵英环,郭贵锁. 基于主题词迭代提取的信息检索算法[J]. 华南理工大学学报(自然科学版),2004(S1):77 - 80.
[16] 俞征鹿,贾佳. 中国科技论文合著情况分析[J]. 全球科技经济瞭望, 2017,32(Z1):92 - 100.
[17] 邹鼎杰. 图情学 4 种两栖类核心期刊合著现象分析[J]. 农业图书情报学刊, 2016,28(3):61 - 64.
[18] 崔林蔚,陆颖. 基于作者署名排序的作者贡献要素分析——以《图书情报工作》2015 - 2016 年作者贡献声明为例[J]. 图书情报工作,2017,61(9):80 - 86.
[19] 左菊. 科研合著中署名顺序与作者贡献研究[D]. 重庆:西南大学,2014.
[20] 贾贤,王霞,李忠富,等. 科技论文中等同贡献作者和共同通讯作者的署名问题[J]. 中国科技期刊研究,2012,23(4):603 - 605.
[21] 李宗红. 利用发文量和被引量综合测评期刊核心著者——以《农业图书情报学刊》为例[J]. 农业图书情报学刊,2007,19(10):161 - 163.
[22] 李品,周金元,杨国立. 基于 CSSCI 的《情报学报》载文被引分析及研究[J]. 图书情报研究,2009,2(2):48 - 52.
[23] 谢瑞霞,李秀霞,韩霞,等. 基于加权被引频次与署名顺序的作者影响力评价指标构建[J]. 情报科学,2018,36(8):90 - 93,111.
[24] MIHALCEA R, TARAU P. TextRank: bringing order into texts [C]// Conference on empirical methods in natural language processing. Stroudsburg: ACL, 2004:404 - 411.
[25] 夏天. 词语位置加权 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2013,29(9):30 - 34.
[26] LIU Z, HUANG W, ZHENG Y, et al. Automatic keyphrase extraction via topic decomposition [C]// Conference on empirical methods in natural language processing. Stroudsburg: ACL, 2010:

366 – 376.
[27] 刘萍,周梦欢. 基于共词网络的专家专长挖掘[J]. 情报科学, 2012,30(12):1815 – 1819.
[28] TSAI T M, SHIH C C, PENG T C, et al. Explore the possibility of utilizing blog semantic analysis for domain expert and opinion mining[C]// Conference on intelligent networking and collaborative systems. Piscataway: IEEE, 2009:241 – 244.
[29] 张晓娟,陆伟,程齐凯. PLSA 在图情领域专家专长识别中的应用[J]. 现代图书情报技术, 2012,28(2):76 – 81.
[30] 杜雨萌,张伟男,刘挺. 基于主题增强卷积神经网络的用户兴趣识别[J]. 计算机研究与发展,2018,55(1):188 – 197.
[31] 宁建飞,刘降珍. 融合 Word2vec 与 TextRank 的关键词抽取研究[J]. 现代图书情报技术, 2016,32(6):20 – 27.
[32] 陆伟,刘杰,秦喜艳. 基于专长词表的图情领域专家检索与评价[J]. 中国图书馆学报, 2010,36(2):70 – 76.
[33] 胡月红,刘萍. 基于本体概念的专长表示研究[J]. 图书情报工作,2012,56(4):17 – 21,40.
[34] 陆伟,武川. 实体链接研究综述[J]. 情报学报,2015,34(1):105 – 112.
[35] MIHALCEA R. Wikify!: linking documents to encyclopedic knowledge[C]// Conference on information and knowledge management. New York: ACM, 2007:233 – 242.
[36] 罗鹏程. 基于概念层次的文档组织方法研究[D]. 北京:北京师范大学,2014.
[37] FERRAGINA P, SCAIELLA U. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities)[C]// Conference on information and knowledge management. New York: ACM, 2010: 1625 – 1628.
[38] TSAI C, ROTH D. Illinois cross-lingual wikifier: grounding entities in many languages to the English wikipedia[C]// International conference on computational linguistics. New York: ICCL, 2016:146 – 150.
[39] 毛进,李纲. 一种基于 OKM 的研究领域专家图谱构建方法[J]. 图书情报工作, 2014,58(14):34 – 40.
[40] 范晓玉,窦永香,赵捧未,等. 融合多源数据的科研人员画像构建方法研究[J]. 图书情报工作,2018,62(15):31 – 40.
[41] 杜建,张玢,唐小利. 作者学术影响力双重测度探讨:引用影响力和合作影响力之整合[J]. 情报学报,2014,33(4):388 – 395.
[42] 潘启树,吴冲,程建霞. 基于模糊 AHP 理论的科学论文学术价值评审研究[J]. 编辑学报,2001,13(1):16 – 18.

作者贡献说明:
陈翀:确定研究方案、撰写论文;
李楠:研究设计、结果评测、撰写论文;
梁冰:完善研究方案及评测方法;
王晨琳:英文标签识别实验、权重算法设计;
徐曾旭林:中文标签识别实验;
郑婷婷:文献调研。

Identifying Expertise Tags of Scholars by Multiple Features of Academic Publications

Chen Chong[1] Li Nan[1] Liang Bing[2] Wang Chenlin[1] Xu Zengxulin[1] Zheng Tingting[1]
[1] School of Government Management, Beijing Normal University, Beijing 100875
[2] Institute of Scientific and Technical Information of China, Beijing 100038

Abstract: [**Purpose/significance**] Identifying expertise tags of scholars is the most critical task in scholar profiling. Expertise tags contribute to finding peer experts, clustering domain scholars and selecting reviewers. [**Method/process**] This study analyzed related factors on the scholar expertise in academic publications, then constructed a hierarchical analysis model on the weight allocation of the factors. The TextRank algorithm has been used to identify topical terms in Chinese corpus, and the conceptual linking technique in English corpus. The extracted terms, together with the previously analyzed factors have been combined to select the expertise tags of the scholars. In this study, a group of honored scholars of different domains have been selected. Their research expertise information from their resumes have been taken as evaluation benchmark. And the expertise tags extracted from their publications have been compared with the benchmark by human judgment and additional semantic similarity judgment. [**Result/conclusion**] The evaluation shows that the expertise tags of 71.9% scholars are acceptable for Chinese, and 77.2% for English. The experiment proves that the method proposed in this article is pragmatic and may lead to reasonable results. The chief innovation of this study lies in three aspects, Firstly, term extraction approaches that suit to different application conditions have been explored, such as the language of publication and the availability of domain knowledge base. Secondly, multiple features have been combined together to identify the expertise tags of scholars, including the content of publications, the substantial contribution to the publications of the scholars, and the influence to the domain of the publications. Thirdly, a reasonable experimental design and evaluation method is proposed, and the proposed approach has been verified by combining manual scoring and semantic calculation results.

Keywords: scholar profiling expertise tagging analytic hierarchy process term extraction evaluation on expertise tagging